Received: 7 September 2023 | Accepted: 20 December 2023

DOI: 10.1002/aps3.11581

APPLICATION ARTICLE

Applications in Plant Sciences

Hijacking a rapid and scalable metagenomic method reveals subgenome dynamics and evolution in polyploid plants

Gillian Reynolds[1,2] | Brendan Mumey[2] | Veronika Strnadova-Neeley[2] | Jennifer Lachowiec[1]

[1]Plant Sciences and Plant Pathology Department, Montana State University, Bozeman, Montana 59717, USA

[2]Gianforte School of Computing, Montana State University, Bozeman, Montana 59717, USA

Correspondence

Jennifer Lachowiec, 119 Plant Biosciences Building, Montana State University, Bozeman, Montana 59717, USA.
Email: jennifer.lachowiec@montana.edu

This article is part of the special issue "Twice as Nice: New Techniques and Discoveries in Polyploid Biology."

Abstract

Premise: The genomes of polyploid plants archive the evolutionary events leading to their present forms. However, plant polyploid genomes present numerous hurdles to the genome comparison algorithms for classification of polyploid types and exploring genome dynamics.

Methods: Here, the problem of intra- and inter-genome comparison for examining polyploid genomes is reframed as a metagenomic problem, enabling the use of the rapid and scalable MinHashing approach. To determine how types of polyploidy are described by this metagenomic approach, plant genomes were examined from across the polyploid spectrum for both *k*-mer composition and frequency with a range of *k*-mer sizes. In this approach, no subgenome-specific *k*-mers are identified; rather, whole-chromosome *k*-mer subspaces were utilized.

Results: Given chromosome-scale genome assemblies with sufficient subgenome-specific repetitive element content, literature-verified subgenomic and genomic evolutionary relationships were revealed, including distinguishing auto- from allopolyploidy and putative progenitor genome assignment. The sequences responsible were the rapidly evolving landscape of transposable elements. An investigation into the MinHashing parameters revealed that the downsampled *k*-mer space (genomic signatures) produced excellent approximations of sequence similarity. Furthermore, the clustering approach used for comparison of the genomic signatures is scrutinized to ensure applicability of the metagenomics-based method.

Discussion: The easily implementable and highly computationally efficient MinHashing-based sequence comparison strategy enables comparative subgenomics and genomics for large and complex polyploid plant genomes. Such comparisons provide evidence for polyploidy-type subgenomic assignments. In cases where subgenome-specific repeat signal may not be adequate given a chromosomes' global *k*-mer profile, alternative methods that are more specific but more computationally complex outperform this approach.

KEYWORDS

genome, *k*-mers, MinHash sketching, polyploidy, transposable elements

Polyploid plants harbor some of the largest and most complex genomes and, consequently, follow varied patterns of chromosomal inheritance that are critical to unravel for understanding transmittance of traits. Polyploids are often classified as belonging to one of three classes: autopolyploid, allopolyploid, and segmental allopolyploid. Whereas autopolyploidy generally refers to an intra-species genome doubling, allopolyploidy generally results from hybridization of distinct taxa (Spoelhof et al., 2017; Nadon and Jackson, 2020). This subgenomic distinctness is of great importance genetically, given that meiotic

recombination is driven by chromosomal sequence and structural similarity (Scott et al., 2023). As such, autopolyploids can exchange genetic material both intra- and inter-subgenomically, whereas inter-subgenomic exchange is rare, although not impossible, for allopolyploids (Spoelhof et al., 2017; Nadon and Jackson, 2020; Scott et al., 2023). Despite the discrete classification into types, polyploidy is a continuum with a myriad of variations identified (Barker et al., 2016; Mason and Wendel, 2020; Blischak et al., 2023). Notably, segmental allopolyploids exhibit auto- and allopolyploidy-like behavior and inheritance (Stebbins, 1947; Mason and Wendel, 2020; Deb et al., 2023).

Determining the type and degree of polyploidization (auto-, allo-, segmental polyploidy) is challenging. However, typing polyploidization is of broad interest, given that polyploids are found across eukaryotes, including fungi and vertebrates (Van de Peer et al., 2017), in addition to plants. The traditional approach includes examining karyotypes to distinguish bivalents and polyvalents of metaphased cells. However, this approach to typing can be inaccurate as autopolyploids, such as *Vaccinium corymbosum* L. (highbush blueberry) (Qu et al., 1998) and *Solanum tuberosum* L. (potato) (Choudhary et al., 2020), can have chromosomes that form bivalents rather than polyvalents, and multivalent associations can be dynamic throughout meiosis (Jones et al., 1996). Karotyping is now complemented by inferring inheritance patterns of alleles to determine ploidy type (Lloyd and Bomblies, 2016), with genome-wide genotyping providing the best resolution but presenting numerous technical challenges (Gerard et al., 2018).

The engineering of genome sequence–based strategies for polyploid typing offers an alternative and complementary approach to the problem. A small number of dedicated programs, such as PolyCRACKER, SubPhaser, PolyReco, and GenomeScope2.0 with Smudgeplot, and a number of genome-specific subgenomic phasing approaches exist for determining a combination of polyploid typing, subgenomic structure, and genomic characteristics (Gordon et al., 2019; Ranallo-Benavidez et al., 2020; Scalabrin et al., 2020; Lovell et al., 2021; Jia et al., 2022; Wang et al., 2022; Goeckeritz et al., 2023; Jin et al., 2023; Session and Rokhsar, 2023). Often these approaches use complex, multistep processes involving numerous programs that pose computational time and space challenges due to the size and complexity of polyploid genomes. However, a common theme among these approaches is the use of subgenome-specific signatures based on *k*-mer content or frequency, where a *k*-mer is a string of nucleotides of length *k*.

In this work, we take inspiration from these existing approaches by utilizing *k*-mer signatures for polyploid typing, and we seek to minimize the computational resources required for analysis and the complexity of implementation. Analogous to PolyCRACKER (Gordon et al., 2019), we reformulate the subgenome identification problem (separating chromosomes into multiple subgenomes) as a metagenomic problem (separating multiple genomes) to take advantage of the pre-existing metagenomic tool sourmash (Pierce et al., 2019). The program utilizes a MinHash sketching technique to down-sample a given sequence's *k*-mer signature and then perform comparative analysis to other sequences. This approach does not require any pre-selection and extraction of genetic features (e.g., protein-coding genes), and so sequences like assembled chromosomes can simply be provided to sourmash for automatic *k*-mer extraction, MinHashing, and comparison. Furthermore, as an alignment-free sequence-comparison technique, it does not suffer from any of the assumptions traditionally held by alignment strategies regarding genome structure (i.e., collinearity), nor their computationally demanding implementations (Zielezinski et al., 2017; Dewey, 2019). Given that polyploid genomes can be very large (e.g., the 17-Gbp hexaploid wheat) and feature multiple collinearity-violating events, sourmash is space- and time-efficient, making it an excellent choice for polyploid sequence comparison in plants and beyond. We explore the capabilities of sourmash for polyploidy typing and allopolyploid progenitor inference and find strong correspondence with previously published findings. We perform a comprehensive assessment of *k*-mer composition and frequency parameters, and finally, we show that the presence of a high repeat content is not a hindrance for this method, but a rather desirable feature.

METHODS

Data acquisition

Chromosome-scale assemblies of allopolyploid, autopolyploid, and segmental allopolyploid genomes were obtained for 16 polyploid plants. All genomes (Table S1, see Supporting Information) were downloaded from the National Center for Biotechnology Information (NCBI) (O'Leary et al., 2016) with the exception of *Solanum tuberosum*, which was downloaded from spudDB (Hirsch et al., 2014) due to availability. If multiple assemblies were available, the NCBI reference assembly was chosen. If multiple species were available but were considered the same crop with the same polyploid structure (e.g., tetraploid cotton), a single domesticated genome was chosen. Chromosomes were separated from the bulk assembly file using an in-house Python script available from https://github.com/Glfrey/Hijacking_Sourmash. Chromosomes were labeled following their naming conventions as given in their publication except in cases where subgenomic clarity was added or names were shortened for data visualizations. The exception is for *Camelina sativa* (L.) Crantz, for which subgenomic assignments were replaced with newer assignments (Mandáková et al., 2019).

Construction of *k*-mer signatures

Sourmash version 4.0.0 (Pierce et al., 2019) was used to generate MinHash *k*-mer signatures for each chromosome for *k*-mers within *k*-mer range 2–20 and then 21–61 in increments of 10

for the default scale factor of 1000. *k*-mer ranges were chosen based on literature that supports the use of small *k*-mers for metagenomics-based binning methods (Dubinkina et al., 2016; Quince et al., 2017; Sedlar et al., 2017) and discussions of sourmash optimal parameters (Brown et al., 2023).

Sourmash was then used to perform *k*-mer signature comparison, calculating similarity of chromosomal signatures via the Jaccard distance for frequency and cosine similarity for composition. Sourmash plots were then used to visualize similarities via a dendrogram, which we assessed for polyploid type identification and insight into genome evolution.

Strict subgenomic clustering determination

As the interpretation of the number of clusters in a dendrogram can be subjective, the categorization of chromosomes to strictly subgenomically cluster was determined by the number of cuts to the dendrogram equal to the number of expected subgenomes minus one (i.e., one cut is sufficient for a tetraploid, two for a hexaploid, which results in the dendrogram being split two and three ways, respectively). This then results in the separated clusters containing only, and all, chromosomes belonging to each subgenome. If a single or double cut to the hierarchical clustering result (representing tetraploid or hexaploid genomes, respectively) could result in distinct clusters containing the chromosomes of each subgenome, the clustering result was deemed to be subgenomically correct. Otherwise, chromosomes were considered non-subgenomically clustered.

Progenitor clustering

For genomes with progenitors that had chromosome-scale assemblies available, progenitor clustering tests were performed in the same way as the subgenomic clustering investigations. This included *Triticum aestivum* L., its known A and D subgenome progenitors *T. urartu* Thumanjan ex Gandilyan and *Aegilops tauschii* Coss., and its potential B subgenome progenitor *A. speltoides* Tausch (Guan et al., 2020; Li et al., 2022). All *Brassica* genomes were involved in the subgenomic analysis alongside their progenitors, *B. napus* L., *B. juncea* (L.) Czern., *B. carinata* A. Braun, *B. oleracea* L., and *B. rapa* L. (Yim et al., 2022). The peanut genome analysis included *Arachis hypogaea* L. and its progenitors *A. duranensis* Krapov. & W. C. Greg. and *A. ipaensis* Krapov. & W. C. Greg. (Chen et al., 2016; Lu et al., 2018). If the subgenome and their progenitor chromosomes were clustered together such that a single cut could be made to separate each of the progenitors and their donated subgenome from the rest of the clusters, they were considered clustered according to their hybridization history.

Detailed clustering analysis

To investigate the causal sequences behind the clustering results and to assess the suitability of the clustering method implemented by sourmash, a second, more detailed analysis was performed for those sequences with repeat-masked genomic sequences available. The repeat-masked and non-repeat-masked chromosomes were downloaded from Ensembl Plants v52 (Yates et al., 2022) and repeat-masking completeness assessed via a custom Python script that enumerated the number of N (any nucleotide base) in the assembly file that were compared against published repeat values for each genome (https://github.com/GLfrey/Hijacking_Sourmash).

Sourmash version 4.0.0 (Pierce et al., 2019) was used to generate MinHash *k*-mer signatures for all sequences for odd *k*-mers within *k*-mer range 3–61 and with scale factors of 1000, 500, 250, and 125. Sourmash was then used to output the similarity scores between the chromosomes for both *k*-mer frequency and composition.

The resulting pairwise similarity scores were then downloaded and analyzed using R Studio 3.6.1 (RStudio Team, 2020). Hierarchical clustering was performed for all sampled *k*-mer sizes and scale factors using hclust for single, complete, average, and Ward's D linkage schemes. The resulting dendrogram that best fit the underlying similarity matrix was determined using a cophenetic correlation strategy, which assesses how faithfully a dendrogram represents the underlying similarity matrix (Saraçli et al., 2013). Generally, a cophenetic correlation of >0.9 is excellent, while <0.7 is considered very poor (Huff, 2001; Rohlf, 2009).

Final clustering with the appropriate scheme and heatmap construction was performed using ComplexHeatmap (Gu et al., 2016). Cuts equal to those expected from the number of subgenomes were then performed using ComplexHeatmap and a visual check of chromosome membership to expected subgenomic clusters was performed. If the chromosomes were correctly clustered according to their subgenomic origin, the dendrogram cut height was recorded for each cut used to separate the subgenomes (e.g., for a tetraploid this would be a single cut to separate the two subgenomes, whereas for a hexaploid this would be two cuts). Dendrogram cut height is indicative of sequence similarity, with larger heights corresponding to greater sequence dissimilarity. Plots describing the relationship between *k*-mers and dendrogram cut height with cophenetic correlation were constructed using ggplot2 (Wickham, 2009). Given that *T. aestivum* is an allohexaploid for which the hybridization timescales and evolutionary relationships are known, the subgenomic clustering correctness was also assessed by whether subgenomes A and B cluster together with the D subgenome as an outgroup (Levy and Feldman, 2022).

In silico transposable element knock-in

Wheat transposable element (TE) coordinates for assembly IWGSCv2.1 were downloaded from the URGI Plant Bioinformatics Facility (https://urgi.versailles.inra.fr/Platform [downloaded June 2023]). Custom Python scripts (https://github.com/Glfrey/Hijacking_Sourmash) were used to extract

and format the repeat information and then extract the corresponding TE sequences from the wheat genome assembly IWGSCv2.1. TEs were grouped according to their type—either long terminal repeat-type (LTR) or non LTR-type. The LTR group was further divided into the RLC and RLG subgroups and the RLX unclassified set according to their Clari-TE classification (https://github.com/jdaron/CLARI-TE/blob/master/clariTeRep_classification.txt). Sequences were appended ("knocked-in") to the *Camelina sativa* genome assembly as listed in Table S2 via a custom Python script that corresponds to the subgenomic grouping and ordering of the polyploid genome described in Kagale et al. (2014). Subgenomic clustering was determined as described above.

RESULTS

k-mer genome analysis differentiates types of polyploidization

Across plants with validated types of polyploidization, we examined whether the sourmash MinHash sketching approach recapitulated polyploid type by using *k*-mer frequency and composition signatures across a range of *k*-mer values (Table S3). We developed a summary scale to describe the results of clustering subgenomic chromosomes based on sourmash signatures and its agreement with known polyploidy type (Figure 1). Given that *k*-mer frequency (rather than *k*-mer composition) more often reflected subgenomic clustering of chromosomes, all species that exhibited correct subgenomic clustering across the majority (≥50%) of tested *k*-mer sizes for *k*-mer frequency and at least some *k*-mer composition are included in the "Subgenomically clusters" class. Those that fail to strictly subgenomically cluster for one particular parameter are classified as "Subgenomically clusters, but…". Species that showed no strict subgenomic clustering as defined in the Methods (e.g., one consistent outlier chromosome), but that show a degree of subgenomic clustering structure are classified as "No subgenomic clusters, but…", whereas those that show no subgenomic clustering are classified as "No subgenomic clusters".

Overall, allopolyploid genomes (Figure 1) showed a high level of subgenomic chromosomal clustering. Eight out of 12 allopolyploid species exhibited clustering for some or all tested parameters (Figure 1, Table S3). Five out of these eight species exhibited highly robust clustering results across all parameters. Three of the five subgenomically clustering allopolyploid genomes (*B. napus*, *Gossypium tomentosum* Nutt. ex Seem., *Gossypium hirsutum* L.) failed to exhibit any subgenomic clustering for *k*-mer composition analysis, placing them in the "Subgenomically clusters, but…" category. Three of the allopolyploids were placed into the "No subgenomic clusters" category, indicating that the genome signatures alone are not sufficient for ploidy typing.

The sourmash approach also contains additional information about the similarity of chromosome signatures. Multiple genera (namely, *Gossypium* and *Brassica*) exhibited asymmetric similarity within the subgenomes, with one subgenome showing greater intra-subgenomic similarity among its member chromosomes than the other subgenome (Figure 2).

Small *k*-mer size ($k < 7$ for *k*-mer frequency, $k < 11$ for *k*-mer composition) was associated with failure to subgenomically cluster for all genomes. For *k*-mer frequency, this was

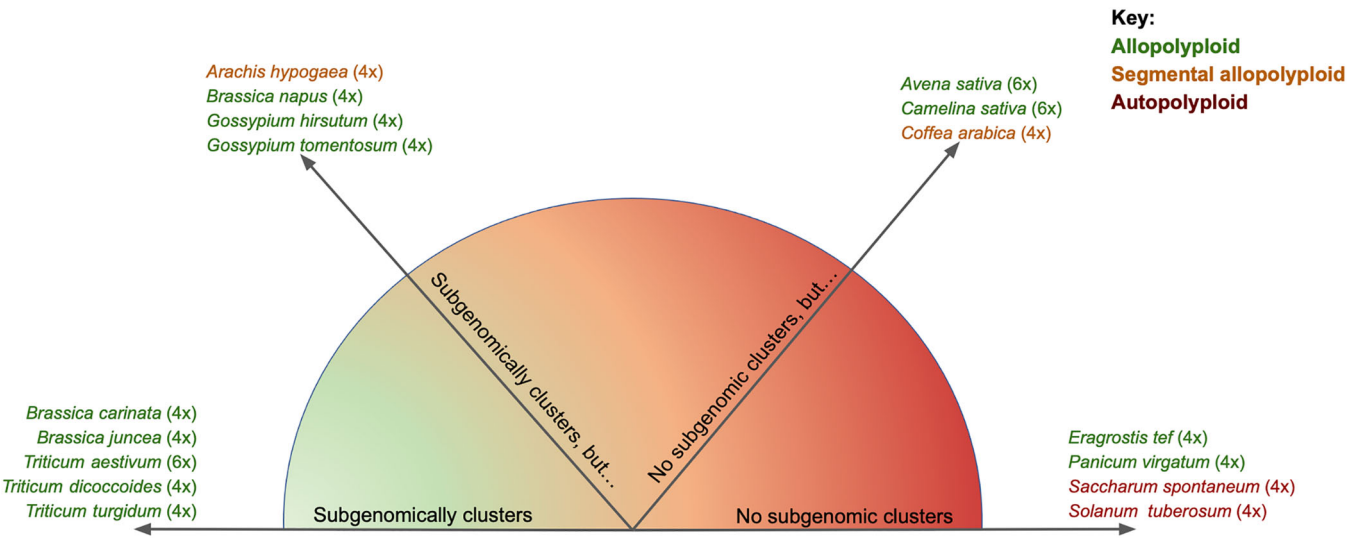

FIGURE 1 The overall tendency of each species, sampled from the three polyploid classes, to subgenomically cluster across a range of sourmash parameters. Species exhibiting subgenomic chromosomal clustering across the majority of parameters belong to the "Subgenomically clusters" class. Those that fail to subgenomically cluster for a particular parameter are classified as "Subgenomically clusters, but…". Species that exhibit a degree of subgenomic clustering structure but are not entirely correct are classified as "No subgenomic clusters, but…", whereas those that show no subgenomic clustering are classified as "No subgenomic clusters". The ploidy level of each species is shown in parentheses.

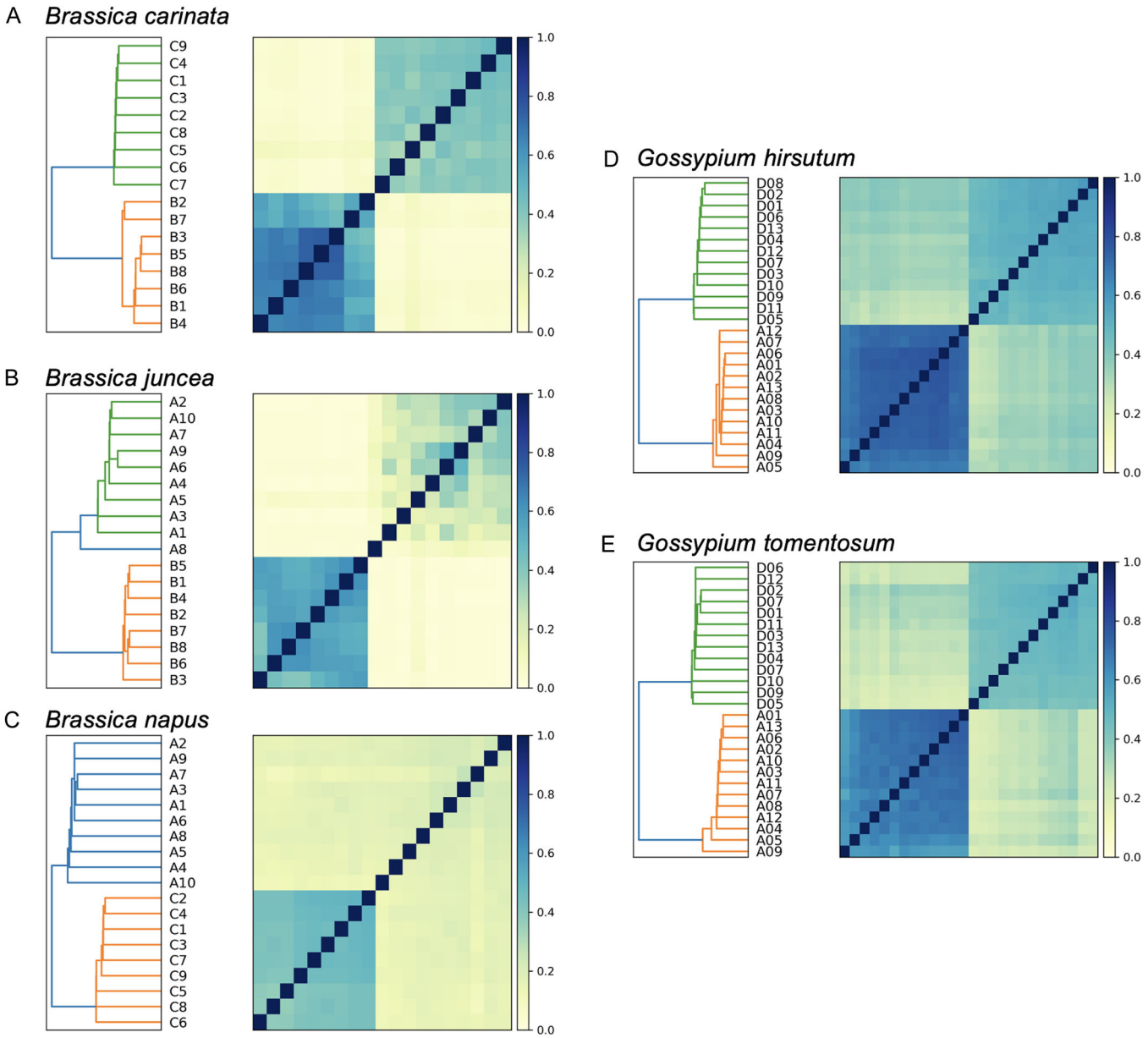

FIGURE 2 Intra-subgenomic chromosomal sequence asymmetries observed for the 21-mer frequency signature. (A) *Brassica carinata*, (B) *B. juncea*, (C) *B. napus*, (D) *Gossypium hirsutum*, and (E) *G. tomentosum* all showed differences in the degree of similarity across chromosomes belonging to each subgenome. Each branch of the dendrograms represents a chromosome from the corresponding genome. The letter labels on the branch tips indicate the subgenome origin, and the number represents the chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale, with dark blue at the value of 1 indicating complete similarity.

due to a lack of any similarity between chromosomes. For *k*-mer composition, there was also a lack of any similarity between chromosomes until $k = 7$, whereafter the similarity between the chromosomes was too high to distinguish subgenomes until $k \geq 11$. Some genomes also showed a failure to cluster at various small *k*-mer sizes ($7 < k < 21$) due to the presence of outgroup chromosomes interrupting subgenomic clustering structure. For example, an exception for *T. aestivum* that failed to recapitulate subgenomic clustering occurred for $k = 7$ due to chromosome 4B being placed as an outgroup chromosome (Figure S1A). Interestingly, neither *T. dicoccoides* (Körn.) Körn. ex Schweinf. nor *T. turgidum* L. showed chromosome 4B acting as an intra-cluster outgroup for $k = 7$ (Figure S1B, S1C). *Brassica juncea* and *B. napus* both showed aberrant results for some small *k*-mer sizes, especially $k < 13$ (Table S3). The information content of short *k*-mers appears inadequate in most cases for subgenomic chromosomal clustering.

Two of the allopolyploid genomes were categorized as "No subgenomic clusters, but…" (Figure 1). *Avena sativa* L. exhibited a strong subgenomic clustering structure for *k*-mer frequency but with consistent intermixing of

chromosomes from the A and D subgenomes (Figure 3A). For *k*-mer composition, a subgenomic-like clustering structure was maintained, with the A and D chromosomes forming homeologous pairs (Figure S2A). The *Camelina sativa* subgenome 3 showed strong inter-chromosomal separation from the other two subgenomes' chromosomes from $k > 9$ onwards for *k*-mer frequency (Figure 3B). For *k*-mer composition, a homeologous-like clustering pattern was maintained for $k > 12$ (Figure S2B).

Several genomes, including the two allopolyploids *Eragrostis tef* (Zuccagni) Trotter and *Panicum virgatum* L. (Figures 3C, D, S2C, D; Table S3) and both tested autopolyploids (*Solanum tuberosum* and *Saccharum spontaneum* L.; Figure 3E, F; Table S3), exhibited no subgenomic clustering structure regardless of parameters. As *P. virgatum* had previously shown subgenomic separation when subgenomic-specific *k*-mers of $k = 15$ were identified (Lovell et al., 2021), sourmash was used to generate signatures and clustering results for $k = 15$ composition and frequency across scale factors. Chromosomes did not subgenomically cluster for the 15-mers regardless of *k*-mer information type and scale factor (Figure S3).

Unlike the allopolyploids, neither of the segmental allopolyploids strictly subgenomically clustered but for different reasons. *Arachis hypogaea* exhibited perfect subgenomic clustering for $k = 11$ and then $k = 14$ onwards (Figure S4A, Table S3), with chromosome 8 interrupting subgenomic clustering by acting as an outgroup otherwise (Figure S4B). For $k < 12$, *Coffea arabica* L. chromosomes showed no coherent pattern of clustering (Table S3). For $k = 13$ onwards, accurate subgenomic clustering structure was consistently interrupted by outgroup chromosomes (Figure 3G, Table S3). Both segmental allopolyploids exhibit a largely homeologous clustering structure for *k*-mer composition (Figure S4C, D).

k-mer analysis clusters polyploid subgenomes with corresponding progenitor genomes

We found that subgenomes accurately cluster with their respective progenitor species genomes, largely following the pattern shown above in which chromosomes subgenomically cluster for *k*-mer frequency and reflect asymmetric rates of genome evolution of subgenomes. For *k*-mer composition, however, chromosomes often clustered into clades containing homeologs.

Triticum species

Modern *T. aestivum* (bread wheat) is allopolyploid (6*x* with A, B, D subgenomes). The chromosomes of *T. aestivum* subgenomically clustered (Table S3), including at $k = 21$ for *k*-mer frequency (Figure 4A) and composition (Figure 4B). The origin of bread wheat includes two subsequent hybridization events between progenitor species. Recent research suggests that these events took place 10,000 and 500,000 years ago (Figure 4C), with the first hybridization event between *T. urartu* and an *A. speltoides* relative and the second between the tetraploid *T. dicoccoides* and *A. tauschii* (Levy and Feldman, 2022). *Triticum aestivum* and its A and D subgenome progenitors, *T. urartu* and *A. speltoides*, respectively, subgenomically clustered throughout the sampled *k*-mer frequency range (Figure 4D, Table S4), except for the short *k*-mers including $k = 7$ and $k = 10$. For $k = 7$, chromosome 4B became an outgroup, whereas the outgroup chromosome was 5D at $k = 10$ (Figure S5A, B). Interestingly, the A and D subgenomes exhibited different clustering patterns with their progenitors for these two *k*-mers. Whereas subgenome A and its progenitor *T. urartu* maintained distinct subgenomic clustering structures, subgenome D and its progenitor exhibit a homeologous clustering structure (Figure S5A, B). In contrast to the *k*-mer frequency used above, both subgenomes and their progenitors exhibited a homeologous clustering structure throughout the sampled *k*-mer range for *k*-mer composition (Figure S5C).

To assess how sourmash would cluster the potential B subgenome progenitor *A. speltoides*, an additional subgenomic clustering investigation was performed with the addition of this genome. *Aegilops speltoides* exhibited a very different relationship to the B subgenome than the A and D progenitors (*T. urartu* and *A. tauschii*) did to their donated subgenomes. Most notably, the A and D subgenomes showed greater subgenomic similarity to their respective progenitors than the B subgenomes showed to *A. speltoides* (Figure S6).

Brassica species

The hybridization between three economically important diploids (*B. nigra* (L.) W. D. J. Koch, *B. rapa*, and *B. oleracea*) formed the three tetraploid crops *B. carinata*, *B. juncea*, and *B. napus* (Xue et al., 2020). Using *k*-mer frequencies, all three tetraploid *Brassica* genomes showed some level of subgenomic–progenitor clustering, and they exhibited varying degrees of sequence similarity between the subgenomes and their progenitor species (Figure S7A–C), which may reflect an unequal evolution of the subgenomes post-hybridization. Similar asymmetries were also observed when assessing these genomes without including progenitors (Figure 2A–C).

Brassica carinata, its B subgenome progenitor *B. nigra*, and its C subgenome progenitor *B. oleracea* formed subgenomic–progenitor clustering pairs for *k*-mer frequency signatures (Figure S7A), with the exception of $k = 11$ and $k = 61$ (Figure S8A, B). As the *k*-mer size increased from $k = 11$ to $k = 61$, the uneven sequence divergence from progenitors became clearer, with the *B. carinata* B subgenome showing greater sequence similarity to its progenitor relative to the C subgenome.

Brassica napus, which also shares the same C subgenome progenitor as *B. oleracea*, showed subgenomic–progenitor clustering at $k = 21$ (Figure S7C), as outgroup chromosomes A01 and A08 interrupted the expected subgenomic–progenitor

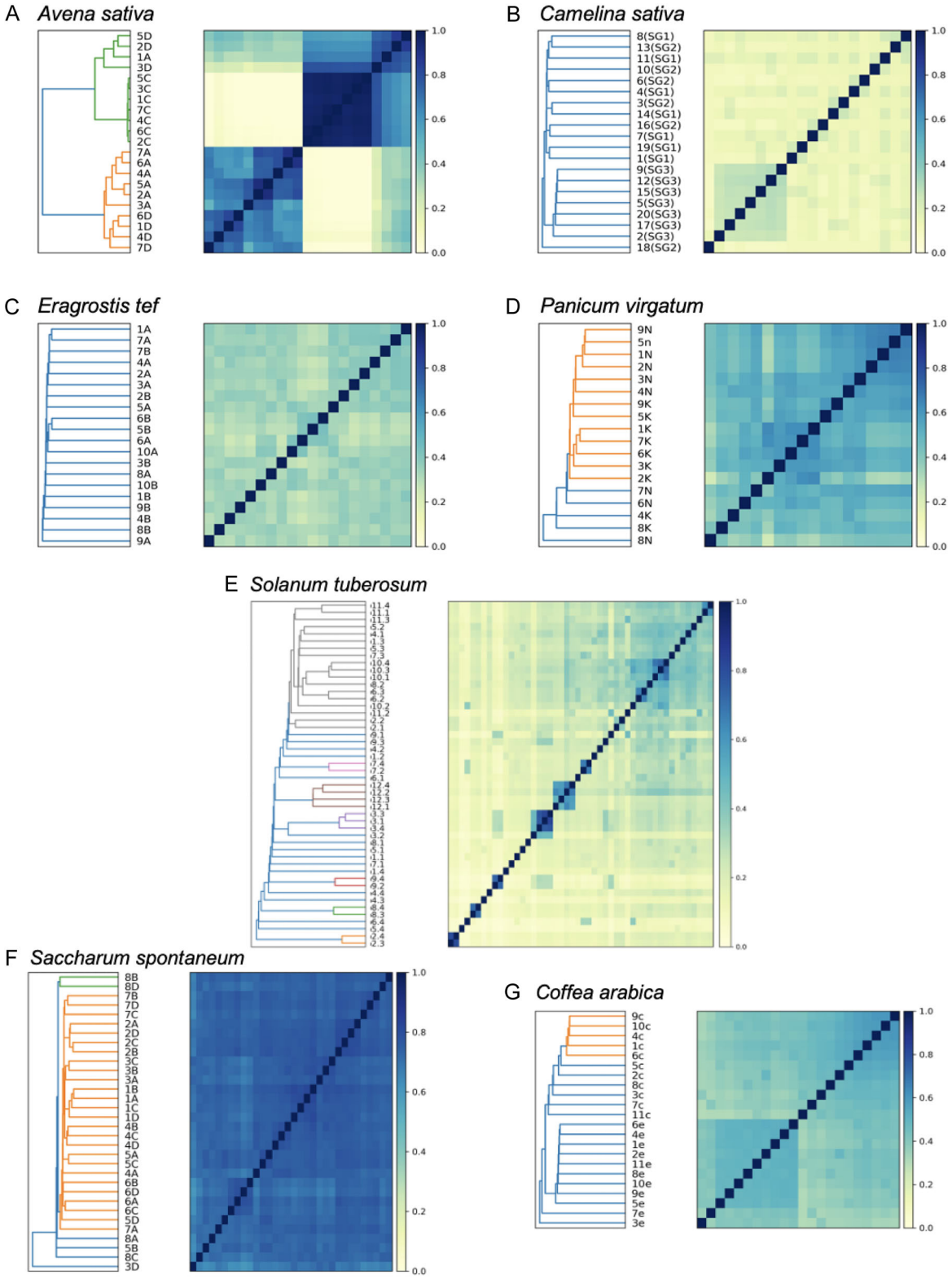

FIGURE 3 Diversity in failure to cluster as demonstrated for 21-mer frequency. (A) *Avena sativa* showed a subgenomic clustering structure interrupted by chromosomal misplacements. (B) *Camelina sativa*, (C) *Eragrostis tef*, and (D) *Panicum virgatum* showed no subgenomic clustering structure. (E) *Solanum tuberosum* and (F) *Saccharum spontaneum* showed no subgenomic clustering structure. (G) *Coffea arabica* subgenomic clustering for 21-mer frequency is shown. Each branch of the dendrograms represents a chromosome from the corresponding genome. The letter labels on the branch tips indicate the subgenome origin (with the exceptions of (B) where subgenome origin is indicated after the label SG with a number or in (E) where subgenome origin is indicated following the decimal point), and the number represents the chromosome number. The heatmap shows the sequence similarity among chromosomes based on *k*-mer signatures using a color scale, with dark blue at the value of 1 indicating complete similarity.

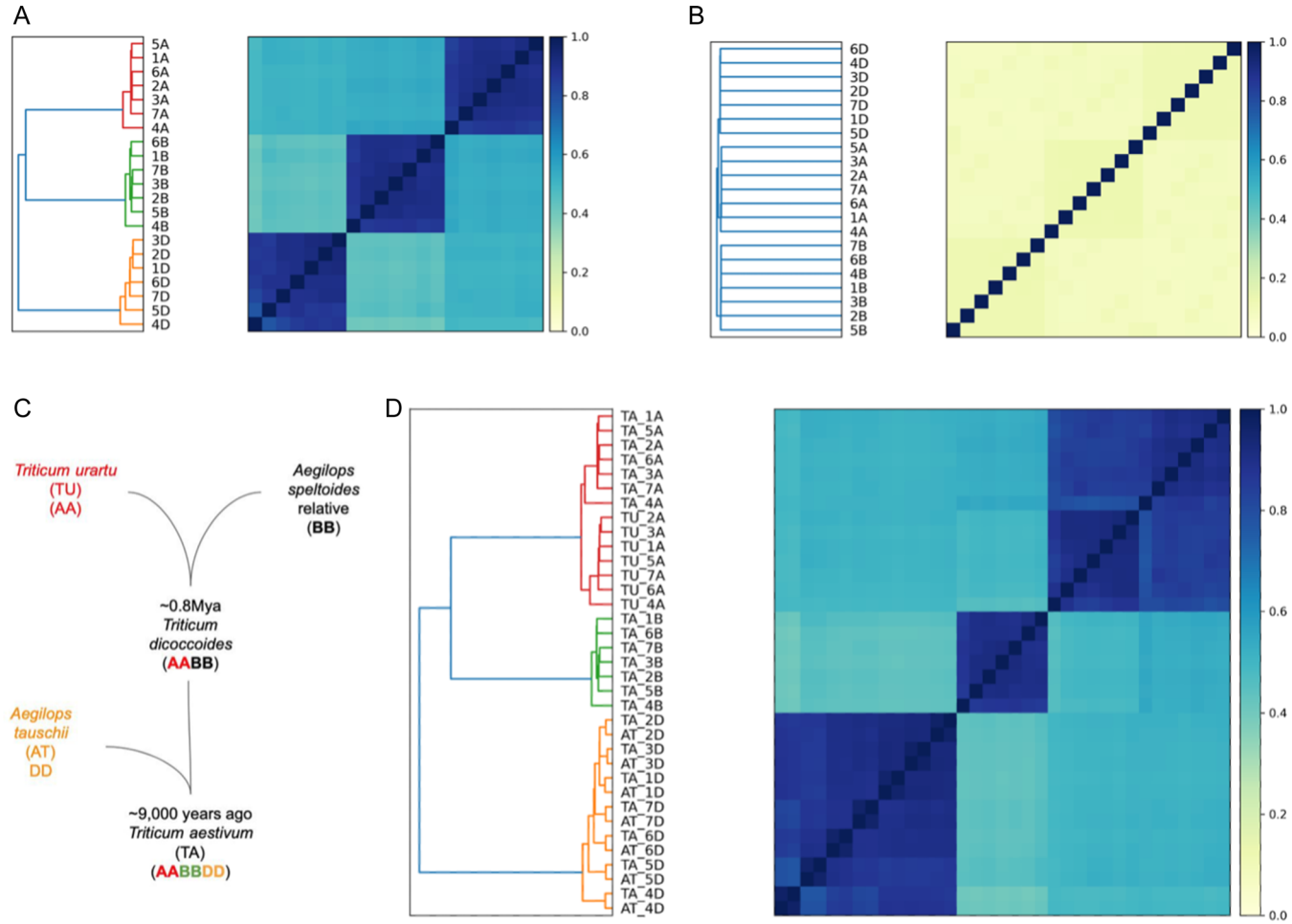

FIGURE 4 *Triticum aestivum* chromosomes subgenomically clustered within genome and with progenitors. Chromosomal subgenomic clustering results for (A) frequency and (B) composition are shown for *T. aestivum* for $k = 21$ using sourmash. (C) Hybridization history for *T. aestivum* from Levy and Feldman (2022) is shown. (D) Subgenomic and progenitor clusters for *T. aestivum* matched models of hybridization history for both the A and D subgenomes found in the same clades with their respective progenitors, *T. urartu* and *Aegilops tauschii*. Each branch of the dendrograms represents a chromosome from the corresponding genome. The letter labels on the branch tips indicate the species origin (TA: *T. aestivum*; TU: *T. urartu*; AT: *A. tauschii*; AS: *A. speltoides*) and subgenome origin (A, B, or D), and the number represents the chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale, with dark blue at the value of 1 indicating complete similarity.

clustering at smaller *k*-mers (Figure S8C). The *B. napus* subgenome C chromosomes showed a greater sequence similarity to their progenitor *B. oleracea* than the subgenome A chromosomes showed to their progenitor *B. rapa*. Conversely, *B. carinata* showed lower subgenome–progenitor similarity for its C subgenome than its B subgenome and progenitor *B. nigra*.

Brassica juncea shares the same B subgenome progenitor as *B. carinata* and also exhibited greater B subgenome–progenitor similarity than observed for its A subgenome and progenitor *B. rapa* (Figure S8). *Brassica napus* also demonstrated this lower A subgenome–progenitor relationship. *Brassica juncea* showed subgenome–progenitor clustering from $k = 7$, but this is interrupted for $k = 8$–9 due to chromosome A01 being an outgroup (as observed for *B. napus* as well), and for $k = 31$ due to the B subgenome clustering away from its progenitor *B. nigra* and instead clustering with the C subgenome and progenitor (Figure S8D, E).

In contrast to these *k*-mer frequency-based results, all *Brassica* genomes did not cluster subgenomically with their progenitor species for *k*-mer composition. Instead, the genomes showed either clear homeologous clustering (*B. napus* and *B. juncea*) or chromosomal pair clustering (*B. carinata*) with their progenitor genomes from either $k = 11$ (*B. napus*) or $k = 21$ (*B. juncea* and *B. carinata*) onward (Figure S9, Table S4).

Arachis

Arachis hypogaea (cultivated peanut) is allotetraploid, arising from hybridization between *A. duranensis* and *A. ipaensis* occurring during domestication (Bertioli et al., 2019). *Arachis hypogaea* showed subgenome–progenitor clustering patterns from $k = 21$ onwards for *k*-mer frequency (Figure S10A, Table S4). For both *A. hypogaea* and its A genome progenitor

A. duranensis, chromosome 08 was an intra-subgenomic outgroup and exhibited lower intra-subgenomic similarity than the other chromosomes. Subgenomic clustering patterns were seen for $k = 9$, but subgenomic clustering is interrupted by the *A. hypogaea* chromosome A8 and *A. duranensis* chromosome 8, which are outgroups (Figure S10B).

Using *k*-mer composition information, *A. hypogaea* chromosomes do not cluster subgenomically with progenitors, similar to the *Brassica* species. Instead, *A. hypogaea* and its progenitors *A. ipaensis* and *A. duranensis* show clear homeologous clustering from $k = 21$ onward for composition (Figure S10C, Table S4), where the chromosomes are organized into a subgenome-like pattern but form no distinct subgenome-specific clusters.

Source of *k*-mer-based subgenome clustering of chromosomes

To understand if a particular class of sequences was driving the subgenomic clustering results, an investigation into the sequences responsible was performed. Given that repetitive elements are rapidly evolving sequences that make up a sizeable portion of most of the genomes under investigation (Table S3), they were naturally under suspicion (Bourque et al., 2018; Session and Rokhsar, 2023). As such, the sourmash procedure was performed with repeat-masked sequences for the *Triticum* genomes.

With the use of repeat-masked sequences, *T. aestivum* and *T. dicoccoides* completely lost subgenomic clustering for both *k*-mer frequency and composition (Figure 5A–D). Where the repeat-rich sequences showed subgenomic clustering (Figures 4A, S1B), the repeat-masked sequences showed homeologous clustering structures for *T. aestivum* and *T. dicoccoides*. The sequences showed less sequence similarity, with barely discernible sequence similarities between the homeologous clusters, in contrast to the very high intra-subgenomic and moderately high inter-subgenomic sequence similarities observed in the presence of repeats. The repeat-masked percentages for the *T. aestivum* and *T. dicoccoides* genome sequences were 86.99% and 87.64%, respectively, which are highly similar to the repeat-masked percentage given in the genome assembly publications (Table S3).

In contrast, *T. turgidum* (Figure 5E) maintained its subgenomic clustering with the loss of repeat sequences for *k*-mer frequency, although there is a marked decrease in sequence similarity for both intra- and inter-subgenomic relationships. The repeat-masked proportion for the *T. turgidum* genome sequence was determined to be 75.45%, which is 6.75% less than the published TE percentage (Table S3). For *k*-mer composition in *T. turgidum*, clustering of homeologs was observed rather than clustering of subgenomes (Figure 5F).

To further investigate the role of repeats in subgenomic clustering, we performed an in silico knock-in of repetitive sequences to non-clustering chromosomes. We selected the non-clustering, allohexapolyploid *C. sativa* as the target because its genome contains only 28% repetitive content (Kagale et al., 2014). In this experiment, annotated TEs, including LTR retrotransposons, were copied from *T. aestivum* (Wicker et al., 2018) and "knocked-in" to the genome of the non-subgenomically clustering allopolyploid *C. sativa* (Table S2). For the knock-in *C. sativa*, only the subgenomic clustering patterns of all TEs and isolated LTR subfamilies mirrored those observed for *T. aestivum* for *k*-mer frequency and composition (Figures 6, S11), although the pattern was markedly weaker for *k*-mer composition than was observed for the *T. aestivum* genome (Figure 4B).

We next examined whether certain LTR families drove subgenomic chromosomal clustering. When the knock-in was limited to specific RLC (Copia) or RLG (Gypsy) LTR sequences, subgenomic clustering remained for *k*-mer frequency, but *k*-mer composition exhibited a homeologous or homeologous-like relationship for RLC and a very weak subgenomic clustering for RLG (Figure S12A–D, Table S5), in contrast to the *T. aestivum* subgenomic relationship (Figure 4B, Table S3). For a knock-in of RLX sequences, which are unclassified LTR retrotransposons, there was a subgenomic-like clustering structure present from 21-mer frequency onwards and a homeologous-like clustering structure for *k*-mer composition (Figure S12E, F; Table S5). For non-LTR TE sequences, we observed a subgenomic clustering structure for the vast majority of *k*-mers tested, but the distinction between and within the subgenomic clusters was weaker than the clustering results for LTR and *T. aestivum* (Figures 4A, S11C–F; Table S5). Interestingly, the *T. aestivum* B subgenome donated a visibly stronger intra-subgenomic similarity for non-LTR sequences throughout the *k*-mer ranges, and for 51-mer frequency, the chromosomes transplanted with 4B and 7B exhibited a distinctly strong similarity (Figure S13). This relationship was shared for *k*-mer composition for small *k*-mers ($k < 15$). Other chromosomes exhibited a homeologous-like structure (Figure S11F) for $k = 15$ onwards (Table S5).

Sourmash parameter suitability

Given that sourmash was developed for metagenomic applications, a comprehensive assessment of parameter suitability for polyploid applications was performed for *T. aestivum*, *T. dicoccoides*, and *T. turgidum*. These species were chosen for their robust subgenomic chromosome clustering across a range of sourmash parameters, which enables an in-depth investigation of parameter interplay. Furthermore, their close relationship with each other and known technical artifacts regarding the *T. turgidum* repeat content allow us to potentially isolate biological and technical factors.

Overall, the parameter with the largest impact on chromosomal subgenomic clustering was whether *k*-mer signatures were based on frequency or composition

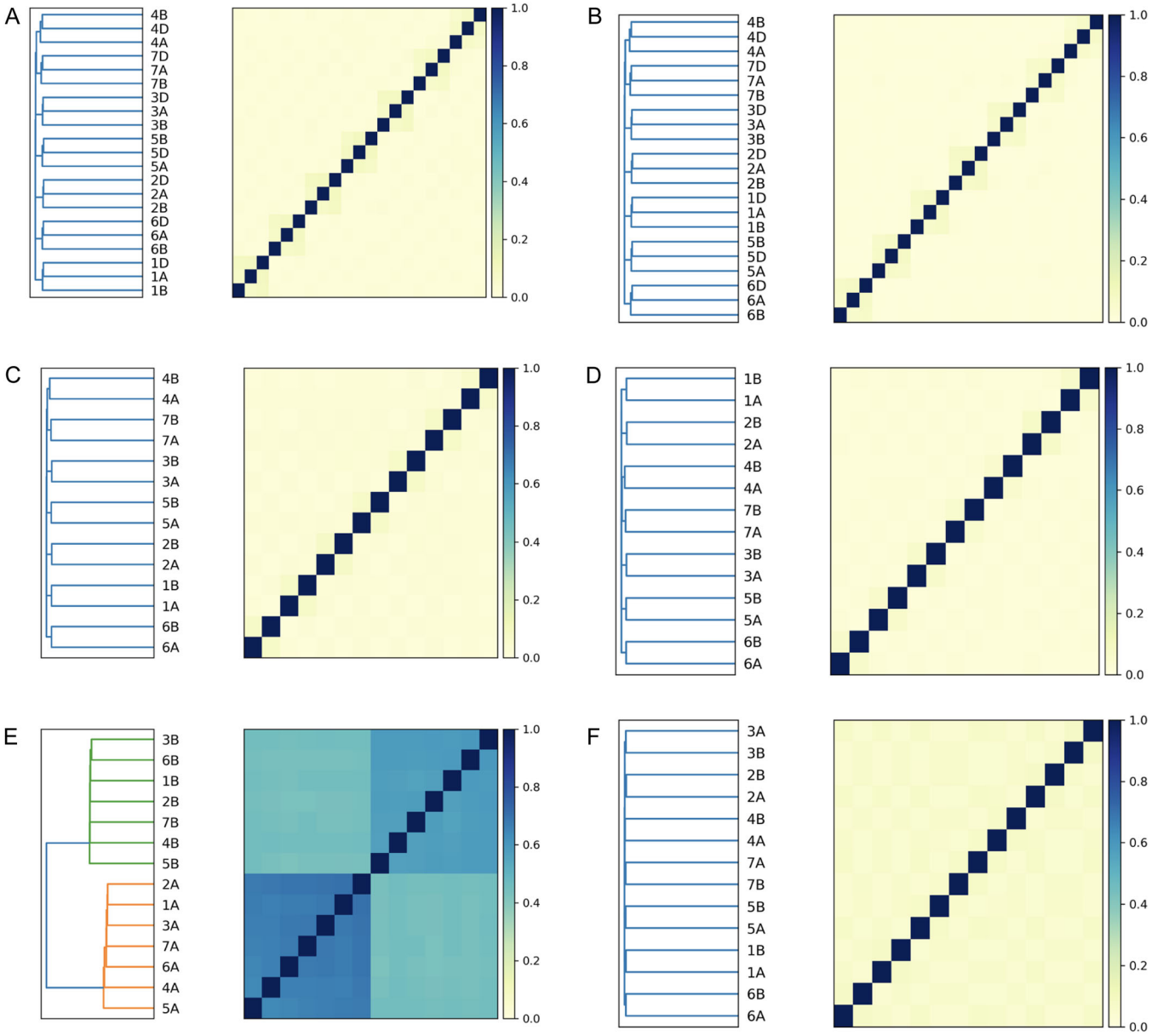

FIGURE 5 Repeat-masked chromosomes fail to subgenomically cluster. *Triticum aestivum* frequency (A) and composition (B), *T. dicoccoides* frequency (C) and composition (D), and *T. turgidum* frequency (E) and composition (F) clustering results are shown. Each branch of the dendrogram represents a chromosome from the genome. The letter labels on the branch tips indicate the subgenome origin, and the number represents the chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale, with dark blue at the value of 1 indicating complete similarity.

(Table S3, Figure S14). The interaction between *k*-mer size and dendrogram cut height demonstrated a positive linear relationship for *k*-mer frequency and composition. However, this relationship was more gradual for *k*-mer frequency, never reaching a value of 1, unlike *k*-mer composition (Figure S14), which showed a cut height of almost 1 (indicating little or no similarity between chromosomes) by $k = 61$.

The scale factor, which controls how much of the *k*-mer space is sampled, had little effect on the subgenomic cut height (Figure S14). This indicates that using the default scale factor of 1000 is just as effective as using small scale factors that require more time and space to work with, especially for genomes with similar characteristics as the *Triticum* genus.

An assessment of the suitability of the hierarchical clustering method implemented in sourmash was performed using the cophenetic correlation, a metric used to assess how faithfully the dendrogram represents the data held in the underlying similarity or dissimilarity matrix (Saraçli et al., 2013). The results revealed that while the single-linkage hierarchical clustering strategy implemented by sourmash is rarely the optimal strategy, there was only a small difference in cophenetic correlations between

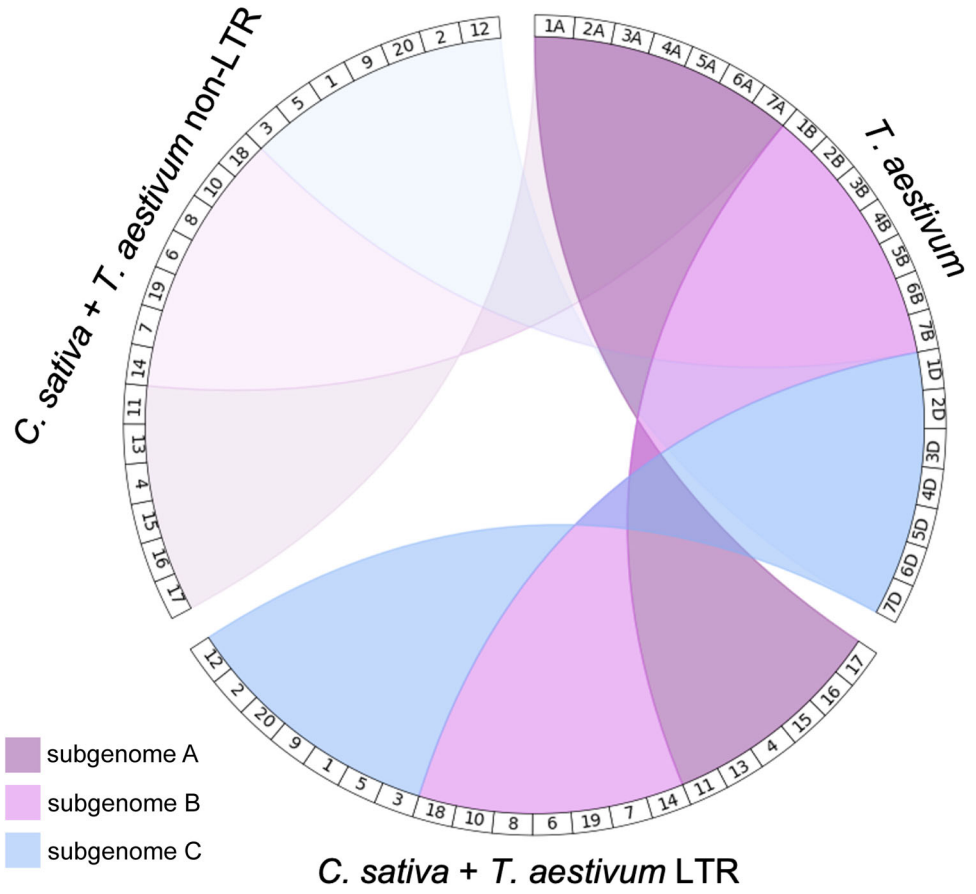

FIGURE 6 Long terminal repeat (LTR) sequences drive chromosome sequence similarities. The sequence similarity relationship across chromosomes (represented with color transparency) is shown for *Camelina sativa* transplanted with *Triticum aestivum* LTR content. While *C. sativa* with the addition of *T. aestivum* LTR sequences maintains a relationship similar to that of the original *T. aestivum* subgenome relationship (Figure 4A), the non-LTR sequences show a marked reduction in sequence similarity, although the subgenomic clustering structure remains.

the often-optimal average-linkage strategy and the implemented single-linkage strategy (Figure S15).

DISCUSSION

The whole-genome MinHash sketching approach for comparative genomics of polyploid crops is capable not only of revealing polyploid type relationships among subgenomes, but also of uncovering evolutionary relationships among species previously described in the literature through sensitivity to the repeat content of the query genomes.

The legacy of the progenitors

We observed that *k*-mer frequency, rather than composition, best recapitulated known polyploidy type. The tendency of chromosomes to subgenomically cluster based on *k*-mer frequency closely matched known ploidy types, with no autopolyploids, half of the segmental allopolyploids, and two-thirds of allopolyploids exhibiting a subgenomic clustering structure for *k*-mer frequency (Figure 1).

Subgenomic distinctness appears to be of great importance for subgenomic clustering here given that meiotic recombination is driven by chromosomal sequence and structural similarity (Scott et al., 2023). For allopolyploids in which subgenomic information is not exchanged on a large enough scale to disrupt a chromosomal signature, the clustering of chromosomes by subgenome is intuitive. In some genomes, such as *T. aestivum* and *B. napus*, the presence of specialized loci prevents meiotic recombination between the subgenomes, thus setting distinct evolutionary trajectories for each subgenome (Le Comber et al., 2010; Spoelhof et al., 2017; Mason and Wendel, 2020).

This pattern of subgenomic clustering across polyploidy types has been observed before by Jia et al. (2022) who, similarly to Gordon et al. (2019), developed a subgenome-specific *k*-mer-based clustering method for polyploid chromosomes. As with sourmash, the method failed on autopolyploids and a number of allopolyploid genomes due to a lack of subgenome-specific *k*-mers. This absence of subgenome-specific *k*-mers is likely caused by the genetic similarity of the subgeneric progenitors. If the initial polyploid genome was created via the hybridization of highly similar genomes, as is expected for autopolyploids

and, to a lesser degree, segmental allopolyploids, then the subgenomes can maintain some degree of inter-subgenomic exchange of genetic information (barring erosion of subgenome-specific repeats over time). This contrasts with allopolyploids, which form through the hybridization of highly distinct subgenomes that are not capable of meiotic transfer of genetic information.

Repeat after me…

Many pieces of evidence support the role of repetitive elements driving subgenomic clustering of chromosomes. TEs and, in particular, the RLC and RLG classes of LTRs are major drivers of subgenomic clustering for our sourmash-based approach in the *T. aestivum* genome. The removal of these sequences resulted in a much weaker subgenomic relationship within and between the subgenomes than is observed when they are present (Figures 5, S11, S12). For *k*-mer frequency, subgenomic clustering of chromosomes was dominated by LTR-type TEs, whereas for *k*-mer composition, other non-TE sequences contributed to the chromosomal signatures. Considering that LTR-type TEs make up the vast majority of plant genomes (Zhou et al., 2021; Jia et al., 2022), their repetitive nature would ensure they dominate any *k*-mer frequency calculations. For *k*-mer composition, however, only *k*-mer presence is recorded, which ensures that *k*-mer contributions from other sequence types—including low-frequency repeats, non-coding sequences, and protein-coding sequences—are equally represented alongside high-copy repeats. A lack of subgenomic clustering for *k*-mer composition but not frequency (as seen for *B. napus* and *Gossypium* species) indicates subgenomic signals originate from subgenome-specific repeat expansions rather than subgenome-specific sequences that would maintain subgenomic clustering in the absence of frequency information (e.g., *Triticum* species).

The dependence of the results on TEs could also explain a number of anomalous results, such as the propensity for the *T. aestivum* A and D subgenomes to cluster together, despite the A and B subgenomes being responsible for the initial hybridization that formed allotetraploid wheat. This can potentially be explained via a B genome–specific TE amplification burst, which appears to have undergone a wave of RLC/Ty1/Copia amplification 1.2 million years ago (Mya) (Avni et al., 2017).

Further evidence for the repeatome-driven clustering results can be found for the subgenomes exhibiting asymmetric subgenome similarities. *Brassica napus*, *B. juncea*, *G. tomentosum*, and *G. hirsutum* are all documented to have subgenomes with asymmetric TE content; these subgenomes have all been reproduced here, with the TE-heavy subgenomes showing greater sequence similarity (Figure 2) (Chalhoub et al., 2014; Sun et al., 2017; Chen et al., 2020; Paritosh et al., 2020).

It is important to note, however, that while LTRs seem to be the drivers of subgenomic clustering for *T. aestivum*, this may not be the case for all species. The *Gossypium* genus, for example, shows a post-hybridization LTR expansion in the D subgenome, which in our work shows less intra-subgenomic similarity than subgenome A (Figure 2D) (Chen et al., 2020). In the African frog *Xenopus laevis*, DNA transposon families distinguished subgenomes via *k*-mer analysis, likely due to their prevalence (Session et al., 2016). Sourmash similarly identified repetitive elements as the sequences responsible for subgenomic clustering.

Where sourmash fails

Unlike alternative *k*-mer-based subgenome-assignment methods, sourmash does not identify subgenome-specific *k*-mers (Jia et al., 2022; Session and Rokhsar, 2023). Instead, it takes a subsample of the whole *k*-mer profile of a given chromosome. Intuitively, the strength of the subgenome-specific *k*-mer signal within that *k*-mer profile will dictate the success of sourmash to cluster the chromosomes subgenomically.

Camelina sativa and *Avena sativa* both failed to subgenomically cluster. For *C. sativa*, subgenome 3 clustered separately from subgenomes 1 and 2, which were often intermingled (Figure 3). This is reflective of the evolutionary history of *C. sativa*, during which two *C. neglecta*-like genomes with distinct chromosome numbers ($n = 6$, $n = 7$) hybridized to form an allotetraploid (subgenome 1 and 2), which was then joined by subgenome 3 donated from *C. hispida* Boiss. (Mandáková et al., 2019). The hybridization between two closely related genomes to form subgenomes 1 and 2 has likely resulted in weak, global, subgenome-specific signals. Session and Rokhsar (2023) were able to take advantage of these signals to facilitate correct subgenomic clustering for all *C. sativa* subgenomes.

There is currently no empirical evidence to explain the failure of *E. tef* and *P. virgatum* to subgenomically cluster using sourmash. For *E. tef*, it is possible that the subgenome-specific repeat signal is weak given that the genome-specific *k*-mers comprise only six out of 65 families of annotated TEs for the genome (VanBuren et al., 2020). Interestingly, centromere-specific repeats rather than subgenome-specific LTRs were used to separate *E. tef* subgenomes (VanBuren et al., 2020). It is also worth noting that the *E. tef* genome has been flagged as "contaminated" on NCBI GenBank (https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_024500355.1/ [accessed 27 March 2024]). For *P. virgatum*, it is possible that the subgenome-specific *k*-mer signal is weak due to time-related erosion of subgenome-specific sequences, given that the hybridization event occurred over 4 Mya (Table S3) (Lovell et al., 2021). Conversely, despite *A. sativa* having formed from a hybridization event more than 7 Mya (Table S3), it did exhibit some subgenomic clustering structure, possibly due to a C subgenome–specific LTR expansion; otherwise, the degradation of subgenome-specific signals is thought to be lineage-specific (Kamal et al., 2022; Session and Rokhsar, 2023). However, we note that the majority of the allopolyploid species examined here have

relatively recent hybridization events (<1 Mya; Table S3), and the capability of sourmash to cluster chromosomes for older hybridization events, such as those found in vertebrates (Van de Peer et al., 2017), is unknown.

A low repeat content could be the driver of the strangely clustering chromosome 8 of *A. hypogaea* and *A. duranensis* (Figure S10). These chromosomes have a much lower percentage of repeat content (49.76% and 44.32%, respectively) than the rest of the chromosomes, which ranges from 72.66–77.98% for the *A. hypogaea* A subgenome (with the B subgenome being even higher) and 49.14–56.67% for *A. duranensis* (Bertioli et al., 2016).

In all, it is clear that sourmash requires a strong, subgenome-specific *k*-mer signal in the global *k*-mer space to produce the correct subgenome assignment of chromosomes. This does not present a problem for methods that identify and utilize subgenome-specific *k*-mers, such as SubPhaser (Jia et al., 2022) and the protocol developed by Session and Rokhsar (2023). The drawback to those approaches is their computational complexity and multi-program implementation. Sourmash therefore complements such approaches and may be a sensible first step for the investigation of ploidy type and assignment of chromosomes to a subgenome.

It is important to note that all subgenomic assessments in this work have been performed for chromosome-scale whole-genome data. Given poorly assembled data, which often suffers from collapsed repetitive sequences, sourmash may not produce optimal subgenomic clustering results.

Progenitor clustering

The sourmash approach produced chromosome clustering of multiple allopolyploids with their progenitors that matched phylogenetic-based approaches and insights into genome evolution pre- and post-hybridization. Consistent with the use of *k*-mer frequency for subgenomic clustering within a species, intra-subgenomic monophyly dominated in comparisons of allopolyploid chromosomes with progenitor chromosomes (Figures 4C, S7, S10). The progenitors and derived subgenomes were often found within the same clade.

We speculate that the progenitor–subgenomic clustering patterns can be attributed to species-specific TEs inherited during hybridization (Bourque et al., 2018). Although the *Brassica* species and *A. hypogaea* underwent post-hybridization repeatome alterations, their relatively recent hybridization suggests that they have a TE landscape more similar to their progenitors (Vitte and Panaud, 2005; Bourque et al., 2018; Wicker et al., 2018; Bariah et al., 2020). This progenitor repeatome legacy is also evidenced through the *A. hypogaea* anomalous chromosome 8 and its progenitor sequence, both of which feature an unusually low repeat content, which together comprised an outgroup.

We further assessed the ability of sourmash to resolve phylogenetically confirmed progenitor subgenomes by including the purported subgenome B progenitor *A. speltoides*. Consistent with the phylogenetic methods that determined *A. speltoides* is not the B subgenome progenitor (Li et al., 2022), we found that *A. speltoides* showed a markedly different relationship to the B subgenome than the A and D subgenomes show with their progenitors (Figure S6).

Impacts of MinHashing and clustering parameters

A comprehensive assessment of sourmash MinHashing parameters (*k*-mer size and scale factor) and clustering parameters (linkage strategy) was performed. Ultimately, scale factor had little impact on the detected subgenome dissimilarity for both *k*-mer frequency and composition (Figures S14, S15). While this has only been tested in the *Triticum* genus, it should be tested on a greater range of species to ensure robustness. Larger scale factors generate smaller genomic signatures, which take less time and require fewer resources to compute and compare, thereby keeping computational overhead to a minimum.

For *k*-mer frequency, the most pronounced differences are for the smaller *k*-mer sizes ($k < 17$), after which they become largely identical. For *k*-mer composition, differences in subgenome dissimilarity for the different scale factors are even less pronounced. As such, it is advisable to use larger *k*-mer sizes (around $k = 21$ and larger) to ensure that the results correctly represent the underlying relationships between the data. This again reflects our findings, with several subgenomes exhibiting anomalous results for small *k*-mer sizes. It is important to note that theoretically the memory consumption of sourmash increases linearly with the number of unique *k*-mers in the downsampled *k*-mer space. In practice, this has such a minimal impact on the size of the resulting set of *k*-mers that sourmash memory usage is little affected by *k*-mer size (Brown, 2023). This contrasts with other *k*-mer-based tools, for which, in the worst case, memory usage can increase at a rate of $2k$, where k is the *k*-mer size (Rødland, 2013).

There is a clear, linearly positive relationship between *k*-mer size and subgenome dissimilarity; for *k*-mer frequency, this increases gradually with *k*-mer size, whereas for *k*-mer composition it increases rapidly, peaking at values close to 1 (Figure S14). The reason behind this is that the *k*-mer space increases with *k*-mer size at a rate of 4^k. For example, for $k = 4$, the *k*-mer space is $4^4 = 256$, whereas for $k = 31$ the *k*-mer space is $4^{31} = 4.611686 \times 10^{18}$. As a general rule, the chances of encountering the same *k*-mer in two (or more) distinct sequences will decline as the *k*-mer space increases (Bussi et al., 2021). In reality, those chances will be influenced by factors such as sequence similarity and sequence length, and while it is important to note that no natural genome contains all possible *k*-mers, the concept holds true nonetheless (Bussi et al., 2021).

A final investigation to assess the optimal *k*-mer clustering strategy utilized cophenetic correlation, which is a measure of how faithfully the similarity is represented by

hierarchical clustering (Saraçli et al., 2013). Given that sourmash was developed for microbial genomes, it is sensible to ensure that the method remains robust for plant genomes. We found that while a minor modification in linkage strategy is advisable for the most accurate results, the clusters identified using sourmash's defaults remain faithful representations of the underlying similarity matrix (Figure S15).

Final recommendations and conclusions

This comprehensive investigation into the implementation of MinHash-based *k*-mer analysis of polyploid crop genomes has demonstrated that such a strategy, as implemented via the metagenomic software package sourmash, can reveal evolutionary relationships and genome dynamics that are verified in the literature. Multiple layers of evidence from experiments conducted herein, combined with published research, support the notion that subgenomic and progenitor clustering results are repeatome driven, possibly by LTRs.

An investigation into MinHash sketching parameters has revealed that the use of *k*-mer frequency or composition directly influences which regions of the genome dominate the results, providing two different windows into comparative polyploid genomics. We find that all other parameters (i.e., hierarchical clustering method) have little impact, although these can be tuned for optimal results on polyploid genomes. Overall, the rapid and highly scalable MinHash sketching method, as implemented by sourmash, produces robust and biologically accurate results for comparative genomic analysis for even the largest and most complex allopolyploid crops.

AUTHOR CONTRIBUTIONS

G.R., B.M., V.S.-N., and J.L. conceived the research. G.R. performed the investigations and formal analysis. G.R. and J.L. produced visualizations. G.R. and J.L. wrote the manuscript. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors thank Jamie Sherman, Greg Chorak, Coltran Hophan-Nichols (Montana State University-Bozeman), and C. Titus Brown (University of California-Davis). Research reported in this publication was supported by the Office of Science (BER), U.S. Department of Energy (grant no. DE-SC0021369), and the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM103474. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Computational efforts were performed on the Hyalite and Tempest High Performance Computing Systems, operated and supported by University Information Technology Research Cyberinfrastructure at Montana State University.

DATA AVAILABILITY STATEMENT

No new data was produced during this study; publicly available was used.

ORCID

Jennifer Lachowiec http://orcid.org/0000-0003-2962-6448

REFERENCES

Avni, R., M. Nave, O. Barad, K. Baruch, S. O. Twardziok, H. Gundlach, I. Hale, et al. 2017. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science (New York, N.Y.)* 357: 93–97.

Bariah, I., D. Keidar-Friedman, and K. Kashkush. 2020. Where the wild things are: Transposable elements as drivers of structural and functional variations in the wheat genome. *Frontiers in Plant Science* 11: 585515. https://doi.org/10.3389/fpls.2020.585515

Barker, M. S., N. Arrigo, A. E. Baniaga, Z. Li, and D. A. Levin. 2016. On the relative abundance of autopolyploids and allopolyploids. *New Phytologist* 210: 391–398.

Bertioli, D. J., S. B. Cannon, L. Froenicke, G. Huang, A. D. Farmer, E. K. S. Cannon, X. Liu, et al. 2016. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nature Genetics* 48: 438–446.

Bertioli, D. J., J. Jenkins, J. Clevenger, O. Dudchenko, D. Gao, G. Seijo, S. C. M. Leal-Bertioli, et al. 2019. The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nature Genetics* 51: 877–884.

Blischak, P. D., M. Sajan, M. S. Barker, and R. N. Gutenkunst. 2023. Demographic history inference and the polyploid continuum. *Genetics* 224: iyad107.

Bourque, G., K. H. Burns, M. Gehring, V. Gorbunova, A. Seluanov, M. Hammell, M. Imbeault, et al. 2018. Ten things you should know about transposable elements. *Genome Biology* 19: 199.

Brown, C. T. 2023. Q: How does memory usage of sourmash change with k-mer size? GitHub Website: https://github.com/sourmash-bio/sourmash/issues/2843 [accessed 22 November 2023].

Brown, C. T., L. Irber, and N. T. Pierce-Ward. 2023. Using sourmash: A practical guide. Website: https://sourmash.readthedocs.io/en/latest/using-sourmash-a-guide.html [accessed 22 November 2023].

Bussi, Y., R. Kapon, and Z. Reich. 2021. Large-scale *k*-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. *PLoS ONE* 16: e0258693.

Chalhoub, B., F. Denoeud, S. Liu, I. A. P. Parkin, H. Tang, X. Wang, J. Chiquet, et al. 2014. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* 345: 950–953.

Chen, X., H. Li, M. K. Pandey, Q. Yang, X. Wang, V. Garg, H. Li, et al. 2016. Draft genome of the peanut A-genome progenitor (*Arachis duranensis*) provides insights into geocarpy, oil biosynthesis, and allergens. *Proceedings of the National Academy of Sciences, USA* 113: 6785–6790.

Chen, Z. J., A. Sreedasyam, A. Ando, Q. Song, L. M. De Santiago, A. M. Hulse-Kemp, M. Ding, et al. 2020. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nature Genetics* 52: 525–533.

Choudhary, A., L. Wright, O. Ponce, J. Chen, A. Prashar, E. Sanchez-Moran, Z. Luo, and L. Compton. 2020. Varietal variation and chromosome behaviour during meiosis in *Solanum tuberosum*. *Heredity* 125: 212–226.

Deb, S. K., P. P. Edger, J. C. Pires, and M. R. McKain. 2023. Patterns, mechanisms, and consequences of homoeologous exchange in allopolyploid angiosperms: A genomic and epigenomic perspective. *New Phytologist* 238: 2284–2304.

Dewey, C. N. 2019. Whole-genome alignment. *In* M. Anisimova [ed.], Evolutionary genomics: Statistical and computational methods, 121–147. Springer, New York, New York, USA.

Dubinkina, V. B., D. S. Ischenko, V. I. Ulyantsev, A. V. Tyakht, and D. G. Alexeev. 2016. Assessment of *k*-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics* 17: 38.

Gerard, D., L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens. 2018. Genotyping polyploids from messy sequencing data. *Genetics* 210: 789–807.

Goeckeritz, C. Z., K. E. Rhoades, K. L. Childs, A. F. Iezzoni, R. VanBuren, and C. A. Hollender. 2023. Genome of tetraploid sour cherry (*Prunus cerasus* L.) 'Montmorency' identifies three distinct ancestral *Prunus* genomes. *Horticulture Research* 10: uhad097.

Gordon, S. P., J. J. Levy, and J. P. Vogel. 2019. PolyCRACKER, a robust method for the unsupervised partitioning of polyploid subgenomes by signatures of repetitive DNA evolution. *BMC Genomics* 20: 580.

Gu, Z., R. Eils, and M. Schlesner. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32: 2847–2849.

Guan, J., D. F. Garcia, Y. Zhou, R. Appels, A. Li, and L. Mao. 2020. The battle to sequence the bread wheat genome: A tale of the three kingdoms. *Genomics, Proteomics & Bioinformatics* 18: 221–229.

Hirsch, C. D., J. P. Hamilton, K. L. Childs, J. Cepela, E. Crisovan, B. Vaillancourt, C. N. Hirsch, et al. 2014. Spud DB: A resource for mining sequences, genotypes, and phenotypes to accelerate potato breeding. *The Plant Genome* 7: plantgenome2013.12.0042. https://doi.org/10.3835/plantgenome2013.12.0042

Huff, D. R. 2001. Genetic characterization of heterogeneous plant populations in forage, turf and native grasses. *In* G. Spangenberg [ed.], Molecular breeding of forage crops, 149–160. Springer, Dordrecht, the Netherlands.

Jia, K.-H., Z.-X. Wang, L. Wang, G.-Y. Li, W. Zhang, X.-L. Wang, F.-J. Xu, et al. 2022. SubPhaser: A robust allopolyploid subgenome phasing method based on subgenome-specific *k*-mers. *New Phytologist* 235: 801–809.

Jin, X., H. Du, C. Zhu, H. Wan, F. Liu, J. Ruan, J. P. Mower, and A. Zhu. 2023. Haplotype-resolved genomes of wild octoploid progenitors illuminate genomic diversifications from wild relatives to cultivated strawberry. *Nature Plants* 9: 1252–1266.

Jones, G. H., K. A. Khazanehdari, and B. V. Ford-Lloyd. 1996. Meiosis in the leek (*Allium porrum* L.) revisited. II. Metaphase I observations. *Heredity* 76: 186–191.

Kagale, S., C. Koh, J. Nixon, V. Bollina, W. E. Clarke, R. Tuteja, C. Spillane, et al. 2014. The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. *Nature Communications* 5: 3706.

Kamal, N., N. Tsardakas Renhuldt, J. Bentzer, H. Gundlach, G. Haberer, A. Juhász, T. Lux, et al. 2022. The mosaic oat genome gives insights into a uniquely healthy cereal crop. *Nature* 606: 113–119.

Le Comber, S. C., M. L. Ainouche, A. Kovarik, and A. R. Leitch. 2010. Making a functional diploid: From polysomic to disomic inheritance. *New Phytologist* 186: 113–122.

Levy, A. A., and M. Feldman. 2022. Evolution and origin of bread wheat. *The Plant Cell* 34: 2549–2567.

Li, L.-F., Z.-B. Zhang, Z.-H. Wang, N. Li, Y. Sha, X.-F. Wang, N. Ding, et al. 2022. Genome sequences of five Sitopsis species of *Aegilops* and the origin of polyploid wheat B subgenome. *Molecular Plant* 15: 488–503.

Lloyd, A., and K. Bomblies. 2016. Meiosis in autopolyploid and allopolyploid *Arabidopsis*. *Current Opinion in Plant Biology* 30: 116–122.

Lovell, J. T., A. H. MacQueen, S. Mamidi, J. Bonnette, J. Jenkins, J. D. Napier, A. Sreedasyam, et al. 2021. Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature* 590: 438–444.

Lu, Q., H. Li, Y. Hong, G. Zhang, S. Wen, X. Li, G. Zhou, et al. 2018. Genome sequencing and analysis of the peanut B-genome progenitor (*Arachis ipaensis*). *Frontiers in Plant Science* 9: 00604. https://doi.org/10.3389/fpls.2018.00604

Mandáková, T., M. Pouch, J. R. Brock, I. A. Al-Shehbaz, and M. A. Lysak. 2019. Origin and evolution of diploid and allopolyploid *Camelina* genomes were accompanied by chromosome shattering. *The Plant Cell* 31: 2596–2612.

Mason, A. S., and J. F. Wendel. 2020. Homoeologous exchanges, segmental allopolyploidy, and polyploid genome evolution. *Frontiers in Genetics* 11: 564174. https://doi.org/10.3389/fgene.2020.01014

Nadon, B., and S. Jackson. 2020. The polyploid origins of crop genomes and their implications: A case study in legumes. *In* D. L. Sparks [ed.], Advances in agronomy, 275–313. Academic Press, Cambridge, Massachusetts, USA.

O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, et al. 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* 44: D733–D745.

Paritosh, K., A. K. Pradhan, and D. Pental. 2020. A highly contiguous genome assembly of *Brassica nigra* (BB) and revised nomenclature for the pseudochromosomes. *BMC Genomics* 21: 887.

Pierce, N. T., L. Irber, T. Reiter, P. Brooks, and C. T. Brown. 2019. Large-scale sequence comparisons with *sourmash*. *F1000 Research* 8: 1006.

Qu, L., J. F. Hancock, and J. H. Whallon. 1998. Evolution in an autopolyploid group displaying predominantly bivalent pairing at meiosis: Genomic similarity of diploid *Vaccinium darrowi* and autotetraploid *V. corymbosum* (Ericaceae). *American Journal of Botany* 85: 698–703.

Quince, C., A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata. 2017. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology* 35: 833–844.

Ranallo-Benavidez, T. R., K. S. Jaron, and M. C. Schatz. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* 11: 1432.

Rødland, E. A. 2013. Compact representation of *k*-mer de Bruijn graphs for genome read assembly. *BMC Bioinformatics* 14: 313.

Rohlf, F. J. 2009. NTSYSpc-Numeric taxonomy and multivariate analysis systems: Getting started guide. Applied Biostatistics Inc., Port Jefferson, New York, USA.

RStudio Team. 2020. RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, Massachusetts, USA. Website: http://www.rstudio.com [accessed 29 February 2024].

Saraçli, S., N. Doğan, and İ. Doğan. 2013. Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications* 2013: 203.

Scalabrin, S., L. Toniutti, G. Di Gaspero, D. Scaglione, G. Magris, M. Vidotto, S. Pinosio, et al. 2020. A single polyploidization event at the origin of the tetraploid genome of *Coffea arabica* is responsible for the extremely low genetic variation in wild and cultivated germplasm. *Scientific Reports* 10: 4642.

Scott, A. D., J. D. Van de Velde, and P. Y. Novikova. 2023. Inference of polyploid origin and inheritance mode from population genomic data. *In* Y. Van de Peer [ed.], Polyploidy: Methods and protocols, 279–295. Springer, New York, New York, USA.

Sedlar, K., K. Kupkova, and I. Provaznik. 2017. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Computational and Structural Biotechnology Journal* 15: 48–55.

Session, A. M., and D. S. Rokhsar. 2023. Transposon signatures of allopolyploid genome evolution. *Nature Communications* 14: 3180.

Session, A. M., Y. Uno, T. Kwon, J. A. Chapman, A. Toyoda, S. Takahashi, A. Fukui, et al. 2016. Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* 538: 336–343.

Spoelhof, J. P., P. S. Soltis, and D. E. Soltis. 2017. Pure polyploidy: Closing the gaps in autopolyploid research. *Journal of Systematics and Evolution* 55: 340–352.

Stebbins, G. L. 1947. Types of polyploids: Their classification and significance. *In* M. Demerec [ed.], Advances in genetics, 403–429. Academic Press, Cambridge, Massachusetts, USA.

Sun, F., G. Fan, Q. Hu, Y. Zhou, M. Guan, C. Tong, J. Li, et al. 2017. The high-quality genome of *Brassica napus* cultivar 'ZS11' reveals the introgression history in semi-winter morphotype. *The Plant Journal* 92: 452–468.

VanBuren, R., C. Man Wai, X. Wang, J. Pardo, A. E. Yocca, H. Wang, S. R. Chaluvadi, et al. 2020. Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff. *Nature Communications* 11: 884.

Van de Peer, Y., E. Mizrachi, and K. Marchal. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* 18: 411–424.

Vitte, C., and O. Panaud. 2005. LTR retrotransposons and flowering plant genome size: Emergence of the increase/decrease model. *Cytogenetic and Genome Research* 110: 91–107.

Wang, F., K. Zhang, R. Zhang, H. Liu, W. Zhang, Z. Jia, and C. Wang. 2022. PolyReco: A method to automatically label collinear regions and recognize polyploidy events based on the KS dotplot. *Frontiers in Genetics* 13: 842387. https://doi.org/10.3389/fgene.2022.842387

Wicker, T., H. Gundlach, M. Spannagl, C. Uauy, P. Borrill, R. H. Ramírez-González, R. De Oliveira, et al. 2018. Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biology* 19: 103.

Wickham, H. 2009. ggplot: Elegant graphics for data analysis. Springer, New York, New York, USA.

Xue, J.-Y., Y. Wang, M. Chen, S. Dong, Z.-Q. Shao, and Y. Liu. 2020. Maternal inheritance of U's triangle and evolutionary process of *Brassica* mitochondrial genomes. *Frontiers in Plant Science* 11: 805. https://doi.org/10.3389/fpls.2020.00805

Yates, A. D., J. Allen, R. M. Amode, A. G. Azov, M. Barba, A. Becerra, J. Bhai, et al. 2022. Ensembl Genomes 2022: An expanding genome resource for non-vertebrates. *Nucleic Acids Research* 50: D996–D1003.

Yim, W. C., M. L. Swain, D. Ma, H. An, K. A. Bird, D. D. Curdie, S. Wang, et al. 2022. The final piece of the Triangle of U: Evolution of the tetraploid *Brassica carinata* genome. *The Plant Cell* 34: 4143–4172.

Zhou, S.-S., X.-M. Yan, K.-F. Zhang, H. Liu, J. Xu, S. Nie, K.-H. Jia, et al. 2021. A comprehensive annotation dataset of intact LTR retrotransposons of 300 plant genomes. *Scientific Data* 8: 174.

Zielezinski, A., S. Vinga, J. Almeida, and W. M. Karlowski. 2017. Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biology* 18: 186.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Figure S1. 7-mer frequency for (A) *Triticum aestivum*, (B) *T. dicoccoides*, and (C) *T. turgidum*. Each branch of the dendrogram represents a chromosome from the corresponding genome. The letter labels on the branch tips indicate the subgenome origin, and the number represents the chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale with dark blue at the value of 1 indicating complete similarity.

Figure S2. Diversity in failure to cluster as demonstrated for 21-mer composition for (A) *Avena sativa*, (B) *Camelina sativa*, (C) *Eragrostis tef*, and (D) *Panicum virgatum*. *Avena sativa* shows a subgenomic clustering structure for the A and D subgenomes, with the B subgenome acting as an outgroup. *Eragrostis tef*, *P. virgatum*, and *Coffea arabica* show homeologous clustering while *C. sativa* shows a homeologous-like clustering structure. The letter labels on the branch tips indicate the subgenome origin (with the exception of (B) where subgenome origin is indicated after the label SG with a number), and the number represents the chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale with dark blue at the value of 1 indicating complete similarity.

Figure S3. 15-mer plots for *Panicum virgatum* for scale factor 1000 frequency (A) and composition (B), scale factor 500 frequency (C) and composition (D), scale factor 250 frequency (E) and composition (F), scale factor 150 frequency (G) and composition (H), and scale factor 50 frequency (I) and composition (J). Each branch of the dendrogram represents a chromosome from the genome. The letter labels on the branch tips indicate the subgenome origin, and the number represents the chromosome number. The heatmap shows the sequence similarity among chromosomes using a color scale with dark blue at the value of 1 indicating complete similarity.

Figure S4. (A) 21-mer frequency results for *Arachis hypogaea*. (B) 13-mer frequency results for *A. hypogaea* where chromosome 08 acts as an outgroup and exhibits much lower similarity to the rest of the subgenome A chromosomes. (C) 21-mer composition results for *A. hypogaea*. (D) 21-mer composition results for *Coffea arabica*. The letter labels on the branch tips indicate the subgenome origin (with the exceptions of (A–C) where subgenome origin is indicated after the label SG with a number), and the number represents the chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale with dark blue at the value of 1 indicating complete similarity.

Figure S5. (A) 7-mer frequency for *Triticum aestivum*, A subgenome progenitor *T. urartu*, and D subgenome progenitor *Aegilops tauschii* shows chromosome 4B as an outgroup. (B) 10-mer frequency for *T. aestivum*, A subgenome progenitor *T. urartu*, and D subgenome progenitor *A. tauschii* shows chromosome 5D as an outgroup. (C) 21-mer composition for *T. aestivum*, A subgenome progenitor *T. urartu*, and D subgenome progenitor *A. tauschii*. Both relevant subgenomes and progenitors exhibit a homeologous clustering structure. Each branch of the dendrograms represents a chromosome from the corresponding genome. The letter labels on the branch tips indicate the species origin (TA: *T. aestivum*; TU: *T. urartu*; AT: *A. tauschii*) and subgenome origin (A, B, or D), and the number represents the chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale with dark blue at the value of 1 indicating complete similarity.

Figure S6. 21-mer (A) frequency and (B) composition for *Triticum aestivum*: A subgenome progenitor *T. urartu*, D subgenome progenitor *Aegilops tauschii*, and potential B subgenome progenitor *A. speltoides*. Both dendrograms and heatmaps show that *A. speltoides* exhibits distinctly less sequence similarity to the *T. aestivum* B subgenome than either A or B progenitors do to their respective subgenomes. Each branch of the dendrograms represents a chromosome from the corresponding genome. The letter labels on the branch tips indicate the species origin (TA: *T. aestivum*; TU: *T. urartu*; AT: *A. tauschii*; AS: *A. speltoides*) and subgenome

origin (A, B, or D), and the number represents the chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale with dark blue at the value of 1 indicating complete similarity.

Figure S7. Asymmetry summary for 21-mer frequency showing progenitor asymmetry in sequence similarity. (A) *Brassica carinata* and its progenitors *B. nigra* and *B. oleracea* and (B) *B. juncea* and its progenitors *B. nigra* and *B. rapa* show greater subgenomic similarity within the B subgenome. (C) *Brassica napus* and its progenitors *B. oleracea* and *B. rapa* show greater subgenomic similarity within the C subgenome. Each branch of the dendrograms represents a chromosome from the corresponding genome. The letter labels on the branch tips indicate the species origin (BN: *B. nigra*; BR: *B. rapa*; BO: *B. oleracea*) and subgenome origin (A, B, or C), and the number represents the chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale with dark blue at the value of 1 indicating complete similarity.

Figure S8. (A, B) *Brassica carinata*, its B genome progenitor *B. nigra*, and its C genome progenitor *B. oleracea* for (A) 11-mer and (B) 61-mer frequency. (C) *Brassica napus* and progenitors *B. oleracea* and *B. rapa* for 9-mer frequency. (D, E) *Brassica juncea* and progenitors *B. nigra* and *B. rapa* for (D) 9-mer and (E) 31-mer frequency. All show non-subgenomic clustering results. Each branch of the dendrograms represents a chromosome from the corresponding genome. The letter labels on the branch tips indicate the species origin (BN: *B. nigra*; BR: *B. rapa*; BO: *B. oleracea*) and subgenome origin (A, B, or C), and the number represents the chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale with dark blue at the value of 1 indicating complete similarity.

Figure S9. (A) *Brassica carinata*, its B genome progenitor *B. nigra*, and its C genome progenitor *B. oleracea*. (B) *Brassica juncea* and progenitors *B. nigra* and *B. rapa*. (C) *Brassica napus* and progenitors *B. oleracea* and *B. rapa*. All show homeologous clustering for 21-mer composition. Each branch of the dendrograms represents a chromosome from the corresponding genome. The letter labels on the branch tips indicate the species origin (BN: *B. nigra*; BR: *B. rapa*; BO: *B. oleracea*) and subgenome origin (A, B, or C), and the number represents the chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale with dark blue at the value of 1 indicating complete similarity.

Figure S10. (A–C) *Arachis hypogaea* and progenitors *A. ipaensis* and *A. duranensis* for (A) 21-mer frequency, (B) 9-mer frequency, and (C) 21-mer composition. (A) Results show correct subgenomic clustering. (B) Results show chromosome 8 from *A. hypogaea* and *A. duranensis* outgrouping. (C) Results show the homeologous clustering structure exhibited for *A. hypogaea* and progenitors for 21-mer onwards for *k*-mer composition. Each branch of the dendrograms represents a chromosome from the corresponding genome. The letter labels on the branch tips indicate the species origin (AD: *A. duranensis*; AH: *A. hypogaea*; AI: *A. ipaensis*) and subgenome origin (SG A, B, or C), and the number represents the chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale with dark blue at the value of 1 indicating complete similarity.

Figure S11. *Camelina sativa* with the addition of transposable element (TE) content for all (A, B) TE 21-mer (A) frequency and (B) composition; (C, D) LTR 21-mer (C) frequency and (D) composition; and (E, F) non-LTR (E) frequency and (F) composition. All show a subgenomic repeat clustering structure but with differences. Both A and C show a very strong intra- and strong inter-subgenomic similarity, although C shows a slight reduction in sequence similarity. E shows a markedly reduced sequence similarity and inter-subgenome similarity has become nonuniform, with subgenome B showing the greatest similarity followed by a D and then A. Inter-subgenome similarity is also reduced and no longer uniform. Both B and D show a subgenomic relationship with little-to-no sequence similarity, while F shows a homeologous clustering structure. The *C. sativa* chromosome numbers are indicated after "chr". The letter labels on the branch tips indicate the subgenome origin, and the number represents the *T. aestivum* chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale with dark blue at the value of 1 indicating complete similarity.

Figure S12. *Triticum aestivum* differing LTR-subclasses of sequence transplanted onto *Camelina sativa*. (A, B) RLC 21-mer (A) frequency and (B) and composition; (C, D) RLG 21-mer (C) frequency and (D) composition; (E, F) RLX (E) frequency and (F) composition. RLC and RLG show subgenomic clustering for *k*-mer frequency (A and C). RLC shows homeologous clustering for *k*-mer composition, while RLG shows homeologous clustering. RLX shows a subgenomic-like clustering structure for *k*-mer frequency, while *k*-mer composition shows a homeologous-like clustering structure, both of which are interrupted by chromosomal misplacements. The *C. sativa* chromosome numbers are indicated after "chr". The letter labels on the branch tips indicate the subgenome origin, and the number represents the *T. aestivum* chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale with dark blue at the value of 1 indicating complete similarity.

Figure S13. *Triticum aestivum* non-LTR sequence transplantation to *Camelina sativa* for (A) 51-mer frequency and (B) 13-mer composition, both of which exhibit a markedly strong relationship for the chromosomes transplanted with *T. aestivum* 4 and 7B sequences. The *C. sativa* chromosome

numbers are indicated after "chr". The letter labels on the branch tips indicate the subgenome origin, and the number represents the *T. aestivum* chromosome number. The heatmap shows the sequence similarity based on *k*-mer signatures among chromosomes using a color scale with dark blue at the value of 1 indicating complete similarity.

Figure S14. Subgenomic cut height for (A) *Triticum aestivum*, (B) *T. dicoccoides*, and (C) *T. turgidum* and its relationship to *k*-mer size for *k*-mer composition and frequency across scale factors 1000, 500, 250, and 125. All three plots show a positive linear relationship for cut height and *k*-mer size, although the relationship for *k*-mer frequency is more gradual yet jagged, whereas *k*-mer composition shows a rapid, smooth relationship. (B) shows two anomalies that correspond to a failure to subgenomically cluster for scale factors 125 ($k = 17$) and 500 ($k = 37$).

Figure S15. Cophenetic correlation scores and their relationships to *k*-mer size for *Triticum aestivum k*-mer frequency (A) and composition (B), *T. dicoccoides k*-mer frequency (C) and composition (D), and *T. turgidum k*-mer frequency (E) and composition (F). All *k*-mer frequency plots show a similar relationship between cophenetic correlation and *k*-mer size, with *k*-mer frequency showing a rapid positive relationship that nears 1 by $k = 7$ and remains there except for small perturbations. For *k*-mer composition, the same rapid rise to near 0.9–1 is followed by a sharp drop at $k = 17$, followed by another rapid rise and gradual peak near 1 for all by $k = 21$, which remains in place for $k = 31$–61 except for a small anomalous result at $k = 37$ for *T. dicoccoides*.

Table S1. Genome assemblies examined.

Table S2. Repeat knock-in experiment design.

Table S3. Summary results of sourmash single genome chromosome clustering with strict dendrogram cut requirements and genome repeat content.

Table S4. Summary results of sourmash progenitor genome chromosome clustering with strict dendrogram cut requirements.

Table S5. Summary results of sourmash clustering with strict dendrogram cut requirements for *Camelina sativa* with *Triticum aestivum* transposable elements transplant.

How to cite this article: Reynolds, G., B. Mumey, V. Strnadova-Neeley, and J. Lachowiec. 2024. Hijacking a rapid and scalable metagenomic method reveals subgenome dynamics and evolution in polyploid plants. *Applications in Plant Sciences* 12(4): e11581. https://doi.org/10.1002/aps3.11581